63-4-4

Interim Report
Tech. Publication 20
Contract FAA/BRD-363
Project No. 204-1

Prepared for the **FEDERAL AVIATION AGENCY**
SYSTEMS RESEARCH AND DEVELOPMENT SERVICE

an Evaluation of 2-7-hr Aviation Terminal-Forecasting Techniques

OCTOBER 1962

7044-40

THE TRAVELERS RESEARCH CENTER INC.

Best
Available
Copy

Interim Report
Technical Publication 20
Contract FAA/BRD-363

AN EVALUATION OF 2-7-hr AVIATION TERMINAL-FORECASTING TECHNIQUES

Isadore Enger
Lawrence J. Reed
James E. MacMonegle

October 1962

Project 204-1

This report has been prepared by The Travelers Research Center, Inc., for the Systems Research and Development Service, Federal Aviation Agency, under Contract FAA/BRD-363. The contents of this report reflect the views of the contractor, who is responsible for the facts and accuracy of the data presented herein, and do not necessarily reflect the official views or policy of the FAA.

THE TRAVELERS RESEARCH CENTER, INC.
650 Main Street Hartford 3, Connecticut

The Travelers Research Center, Inc.
650 Main Street, Hartford 3, Connecticut.
AN EVALUATION OF 2-7-hr AVIATION TERMINAL-FORECASTING TECHNIQUES,
Isadore Enger, Lawrence J. Reed, and James E. MacMonegle. October 1962.
38 pp. incl. appendix, 1 illus., 24 tbls., 4 refs.; interim rpt.
(FAA/BRD-363)

ABSTRACT

An evaluation of short-period (2-7-hr) ceiling and visibility terminal forecasting techniques indicates that it is possible to prepare objective forecasts of these critical aviation weather parameters which yield a statistically significant improvement over those presently provided in routine operations.

Two types of forecasts were examined: probability forecasts and categorical forecasts. Probability forecasts were evaluated by means of the Brier-Allen P-score, which measures the sharpness and validity of probabilities. Four objective probability forecast procedures and special subjective probability forecasts were compared. The control technique was a climatological procedure that specifies the climatological probability of the events conditional on the initial condition, the season of the year, and the time of day. This is called climatological expectancy of persistence (CEP). Relative to this technique, the best technique is multiple-discriminant analysis (MDA), which achieves a percentage increase ranging from 7.9 to 12.5 for ceiling and 3.6 to 5.6 for visibility.

Categorical forecasts were judged by the Bryan score, which measures the skill in forecasting operationally important categories of ceiling and visibility. The control forecasting procedure was designated as persistence, which is a specification that the weather will remain unchanged. Forecasts prepared by six procedures, including presently available operational aviation weather forecasts prepared subjectively, were evaluated. Procedures producing probability forecasts were converted to categorical forecasts by the use of a loss function that maximizes the Bryan score.

Relative to the control technique (persistence), the MDA procedure yielded the greatest improvement in Bryan score. This improvement ranged from 9.9 to 14.7% for ceiling and 13.3 to 19.2% for visibility. Presently available subjective aviation forecasts yielded an increment ranging from -3.2 to 14.7% for ceiling and 1.7 to 2.7% for visibility.

Special evaluations of experimental subjective probability forecasts revealed that they were inferior to MDA forecasts in terms of the Brier-Allen P-score; but, when converted to categorical forecasts, they yielded the best Bryan scores of any forecast technique tested, including MDA and the subjectively prepared categorical forecasts available in routine operations.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1.0 INTRODUCTION

The design of the Common Aviation Weather System (CAWS) of the Federal Aviation Agency [2] specifies important requirements for weather-data processing techniques to provide short-period predictions of terminal weather conditions. As part of the weather-data processing development program (Contract FAA/BRD-363) in support of CAWS, The Travelers Research Center has undertaken an extensive program to test and evaluate certain terminal weather forecasting techniques to determine their suitability for use in CAWS. The purpose of this test and evaluation program was to compare the accuracy of objective techniques with one another, and with the terminal forecasts presently available on a routine operational basis that are prepared subjectively by station forecasters.

The design of the Common Aviation Weather System calls for large numbers of terminal forecasts to be prepared with great speed to meet the system output requirements. After a preliminary evaluation of available techniques that might be suitable for this purpose, it was decided to focus attention at this time on the engineering of objective statistical procedures that are readily adaptable to machine computation.

This decision was based largely on two considerations. First, it was felt that in the light of our knowledge of the physical processes governing small-scale phenomena of importance in aviation terminal weather problems, techniques based upon the use of the physical equations would require an extensive and indeterminate developmental period, and would be best considered as a logical follow-up development to the statistical procedures. Second, it was felt that techniques that could provide probabilistic predictions would be of great value in increasing the utility of terminal weather information, permitting as they do the communication of the degree of certainty of the forecasts, and providing a rational basis for aviation operational and planning decisions. This report contains a description of the extensive test program and a summary evaluation of the results.

1.1 History of the Test

Originally, the test and evaluation of various terminal forecasting techniques was specified as a requirement of the weather forecasting technique development work in support of Weather Observing and Forecasting System 433L, Contract AF 30(635)-14459. Subsequently, the Federal Aviation Agency assumed responsibility for certain phases of the technique development work, including this test and evaluation program. A Forecast Evaluation Working Group was formed by the government to specify the manner in which the evaluation was to be conducted. Its membership included representatives of the U.S. Air Force, the Federal Aviation Agency, and the U.S. Weather Bureau, as well as the contractors concerned, The United Aircraft Corporation and The Travelers Research Center, Inc. The Working Group devised a plan specifying the conditions of the test. Among these conditions were: the terminals at which forecasts were to be made; the forecast lengths; the valid times of the forecasts; the elements to be forecast and the

class limits for categorizing these elements; the year for which forecasts would be evaluated; and the method for scoring the forecasts. This plan was submitted to the Federal Aviation Agency as a technical memorandum [5] prepared by The Travelers Research Center, Inc. The plan was accepted by the FAA.

1.2 Scope of the Test

The test was designed to answer two questions:

(a) How do objective statistical forecasts compare with operationally available subjective forecasts?

(b) Which of the statistical techniques tested is most accurate?

After a survey of the available statistical techniques, the data sources, and the time available for the test, a selection of techniques and data was made and approved by the government. Five techniques were agreed upon, and the data to be used were restricted to surface hourly airways observations.

It is important to emphasize that this test only establishes an important and valuable bench mark that enables the assessment of how well we can presently do in terminal forecasting. Future development of statistical techniques using upper-air data, derived parameters, and additional stations and data may well produce improved forecasts. The continuing research in small-scale dynamical weather prediction and the introduction of denser networks of stations may also contribute significantly to our ability to improve terminal forecasting capabilities.

1.3 Organization of the Report

The main body of this report is a summary description of the test, its results, and the conclusions to be drawn from it. Appendix A is a nonmathematical description of the Bryan score. A supplement to the report contains a detailed exposition of the test; a list of all the variables presented to the statistical techniques and all the variables used to produce the statistical forecasts; and all the contingency tables of forecast versus observed values and verification scores. These three sections, together with the computer programs, which are available, contain sufficient information to permit other investigators to duplicate any or all parts of the evaluation.

2.0 DESCRIPTION OF THE TEST PLAN

The execution of the test plan included the following steps:

(a) the specification by the government of the terminals, elements, forecast lengths, limits for categorization, statistical techniques, input and valid times, subjective forecasts, and verification procedures,

(b) the designation by the government of a year, henceforth called the evaluation year, on whose data all techniques would be tested,

(c) the collection and processing of dependent and independent (evaluation-year) data,

(d) the development of the statistical forecast techniques on the dependent data,

(e) the production of statistical forecasts on independent data,

(f) the collection and decoding of subjective forecasts, and

(g) the verification of all forecasts.

2.1 Definition of Forecast Length

Figure 2-1 shows how forecast length is defined.

Input time is the time of observation of the data used in the preparation of a terminal forecast. For statistical forecasts, the input time was the observation time of the data used in making a forecast. For subjective forecasts, such as FTs and TAFORs, the input time is uncertain but was generally assumed to be the whole hour preceding the filing time of the forecast. The filing time is the time by which a forecast must be given to communications personnel to ensure meeting communications schedules. For example,

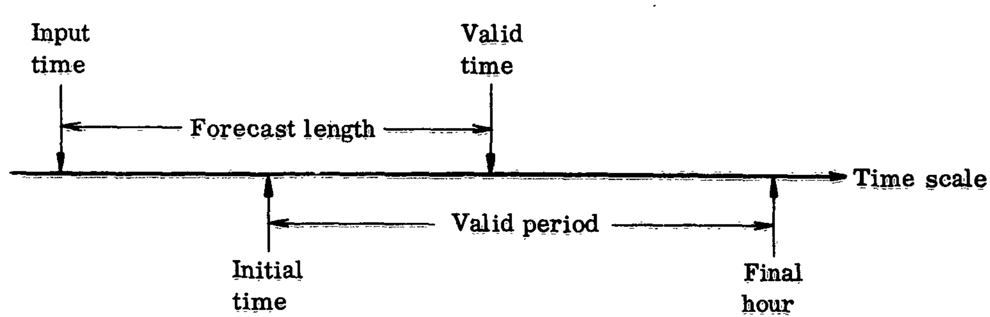

Fig. 2-1. Definition of input time, forecast length, initial time, valid time, and valid period for terminal forecasts.

a forecast with a filing time of 1130 was assumed to have an input time of 1100.

Valid period is the specified period of time during which a forecast is valid. For example, an FT1 forecast is delivered to communications personnel at 1130 and designated as a 12-hr forecast starting at 1200 and ending at 2400. The valid period for this forecast is 1200–2400.

Initial time is the beginning of the first hour of the valid period. In the example cited above, the initial time is 1200.

Final time is the end of the last hour of the valid period. In the example cited above, the final time is 2400.

Valid time is that instant of time at which the forecast value of a parameter is expected to occur. For example, if a 500-ft ceiling is forecast to occur at 1400, the valid time of the forecast is 1400.

Forecast length is the difference between valid time and input time.

2.2 Terminals, Elements, Forecast Lengths, and Limits

The seven terminals approved by the Working Group are given in Table 2-1. The forecast elements are ceiling and visibility. The approved forecast lengths for each station are listed in the table. The Group decided that, for the purpose of verifying the forecasts, each element at each station would be classified into five categories. The limits for categorization are also listed in Table 2-1.

In meteorology, a variable for which a forecast is required is often termed a

TABLE 2-1
TERMINALS, ELEMENTS, FORECAST LENGTHS, AND LIMITS
DESIGNATED BY THE WORKING GROUP FOR THE EVALUATION

Sta*	Forecast lengths, hr	Upper limit of ceiling categories, ft					Upper limit of visibility categories, mi				
		1	2	3	4	5	1	2	3	4	5
ACY	3,5,7	<200	< 500	< 1000	<3000	≤unl	< 0.5	<1	< 2	< 3	≤ unl
CEF	2,3,4,6	<200	< 600	< 1500	<5000	≤unl	< 0.5	<1	< 3	< 5	≤ unl
DCA	2,3,5,7	<200	< 500	< 1000	<3000	≤unl	< 0.5	<1	< 2	< 3	≤ unl
IDL	2,3,5,7	<200	< 500	< 1000	<3000	≤unl	< 0.5	<1	< 2	< 3	≤ unl
OFF	2,4,6	<300	<1000	< 1500	<5000	≤unl	< 0.5	<1	< 2	< 5	≤ unl
RND	2,4,6	<200	< 400	< 1500	<5000	≤unl	< 0.5	<1	< 3	< 5	≤ unl
WRI	2,4,6	<200	< 500	< 1500	<5000	≤unl	< 0.5	<1	< 3	< 5	≤ unl

*ACY Atlantic City Airport
CEF Westover Air Force Base
DCA Washington National Airport
IDL Idlewild International Airport
OFF Offutt Air Force Base
RND Randolph Air Force Base
WRI McGuire Air Force Base

predictand. In this report, a predictand is defined as a specific element at a specific station for a specific forecast length. For example, 2-hr ceiling height at Idlewild is one predictand, 3-hr ceiling height at Idlewild is a second predictand, 3-hr visibility at Idlewild is another predictand, etc. Table 2-1 lists 48 predictands, comprising two elements at seven stations for either three or four forecast lengths.

2.3 Collection and Preprocessing of Data

The data consisted of a dependent set covering the 10 years from 1949 through 1958 and an independent set (evaluation-year data) extending over the year from October 1, 1960, through September 30, 1961. Standard hourly airways observations as punched from WBAN-10A and -10B forms were obtained from Asheville on IBM-705 magnetic tapes. There were approximately 96,000 surface observations for each of 53 stations, where an "observation" contains about 25 elements. Thus, some 125,000,000 pieces of data were processed in carrying out the test.

2.3.1 Dependent Data

Data on the IBM-705 tapes were edited and transferred to IBM-7090 tapes. The data were then processed through a computer program, which rearranged the data from observation to vector format, wherein a single element at a single station forms one record on the tape. The same program adjusted all data to Eastern Standard Time, categorized ceiling heights and visibility, and added constants to certain elements (such as temperature) to ensure that every value was positive, and all data were "packed" so that one machine location contained more than one datum.

Because of restrictions on the size of computer storage and the need for speed of operation, not all data for the complete 24 hours of each day for the 10 years of the dependent sample (a total of 87,000 hours) were used. A random selection of from 6,000 to 10,000 hours was made and used. Gross error checks were made to ensure that, for any hour chosen, all elements in all observations at every station used were correct. Finally, data at various stations were combined to form networks of stations, one network for each of the seven stations. All this preprocessing resulted in individual computer tapes for each of the seven networks.

2.3.2 Independent Data

The independent data were processed similarly. Because these data were in slightly different form, some additional computer programs were written. A set of data to match the form of the dependent data was developed.

2.4 Development of Statistical Forecasting Techniques

The Working Group selected the following statistical techniques to be tested:

TABLE 2-2
STATION NETWORKS USED IN THE DEVELOPMENT
OF TERMINAL FORECASTING TECHNIQUES FOR THE (UNDERLINED) TEST STATIONS

<u>Atlantic City WBAS, N.J.</u>

Lakehurst NAS, N.J.
McGuire AFB, N.J.
Millville, N.J.
Norfolk, Va.
Olmstead AFB, Pa.
Philadelphia, Pa.
Salisbury, Md.
Washington N.A., D.C.
Wilkes Barre/Scranton, Pa.

<u>Idlewild I.A., N.Y.</u>

Albany, N.Y.
Binghamton, N.Y.
Concord, N.H.
Lakehurst NAS, N.J.
Olmstead AFB, Pa.
Providence, R.I.
Salisbury, Md.
Suffolk County AFB, N.Y.
Teterboro, N.J.
Windsor Locks, Conn.

<u>McGuire AFB, N.J.</u>

Albany, N.Y.
Allentown, Pa.
Atlantic City WBAS, N.J.
Lakehurst NAS, N.J.
Newark, N.J.
Norfolk, Va.
Philadelphia, Pa.
Providence, R.I.
Washington N.A., D.C.
Williamsport, Pa.

<u>Offutt AFB, Neb.</u>

Des Moines, Iowa
Grand Island, Neb.
Huron, S.D.
Kansas City, Mo.
Minneapolis, Minn.
Moline, Ill.
North Platte, Neb.
Schilling AFB, Neb.
Sioux Falls, S.D.
Springfield, Mo.

<u>Randolph AFB, Tex.</u>

Bergstrom AFB, Tex.
Brownsville, Tex.
Connally AFB, Tex.
Corpus Christi, Tex.
Ellington AFB, Tex.
Fort Worth, Tex.
Lake Charles, La.
Laredo AFB, Tex.
Laughlin AFB, Tex.

<u>Washington N.A., D.C.</u>

Annapolis NAF, Md.
Atlantic City WBAS, N.J.
Gordonsville, Va.
Martinsburg, W. Va.
Norfolk, Va.
Patuxent River NAS, Md.
Pittsburgh, Pa.
Roanoke, Va.
Williamsport, Pa.

<u>Westover AFB, Mass.</u>

Albany, N.Y.
Atlantic City WBAS, N.J.
Burlington, Vt.
Hanscom Field, Mass.
Idlewild, N.Y.
Portland, Me.
Providence, R.I.
Syracuse, N.Y.
Wilkes Barre/Scranton, Pa.
Windsor Locks, Conn.

(a) persistence,

(b) climatological expectancy of persistence (CEP),

(c) grouping,

(d) Lund contingency prognosis,

(e) multiple-discriminant analysis (MDA), and

(f) the Lewis technique.

Each technique, except persistence, requires the preparation of forecast tables, equations, and/or constants from a dependent sample of data. For each test station, a surrounding network of stations was selected and prepared.

2.4.1 Test Station Networks

The test station networks were selected by a committee of experienced forecasters. The choice was limited to the data available on magnetic tape as received from the National Weather Records Center, with the additional constraint that no network could exceed 11 stations. As closely as the available data permitted, the network stations were selected so that they formed two concentric circles about the predictand station. The inner circle varied from 25 to 100 mi, and the outer varied from 125 to about 250 mi. The station networks are listed in Table 2-2.

2.4.2 Control Techniques

It was important that control techniques be established against which the performance of other techniques would be judged. Separate control techniques for categorical and probability forecasts were designated. Persistence was designated as the control technique for categorical forecasts and CEP as the control technique for probability forecasts.

2.4.2.1 Persistence Forecasts

Persistence forecasts are simple statements that the weather at the valid time will be the same as the weather at the input time. No preparation of forecast tables or equations is required.

2.4.2.2. Climatological-expectancy-of-persistence Forecasts

CEP forecasts are persistence forecasts conditional on the initial conditions, the season of the year, and time of the day. Dependent data consisting of every even hour in the 10 years were stratified into two seasons—May through October and November through April—and then into two diurnal periods. This stratification yielded four sets of data. For each set, a frequency-count table was computed. Table 2-3 is an example. The value 26 in the first row represents the number of times that ceilings below 200 ft at the input time were followed by ceilings below 200 ft 5 hr later. Other entries in the table have similar meaning.

TABLE 2-3
FREQUENCY COUNT OF IDLEWILD CEILINGS*

Ceiling class interval, ft	i	1	2	3	4	5	Total
C < 200	1	26	34	19	16	53	148
200 ≤ C < 500	2	40	120	70	66	93	389
500 ≤ C < 1000	3	23	97	267	175	167	729
1000 ≤ C < 3000	4	6	51	159	384	467	1067
3000 ≤ C	5	22	61	144	421	7886	8534

*5-hr forecast. November-April. Input hours 01E to 13E.

TABLE 2-4
CLIMATOLOGICAL-EXPECTANCY-OF-PERSISTENCE FORECAST TABLE†

Idlewild ceiling class interval, ft	i	1	2	3	4	5
C < 200	1	0.176	0.230	0.128	0.108	0.358
200 ≤ C < 500	2	0.103	0.308	0.180	0.170	0.239
500 ≤ C < 1000	3	0.032	0.133	0.366	0.240	0.229
1000 ≤ C < 3000	4	0.006	0.048	0.149	0.360	0.438
3000 ≤ C	5	0.003	0.007	0.017	0.049	0.924

†5-hr forecast. November-April. Input hours 01E to 13E.

A forecast table, Table 2-4, was obtained by dividing each entry of Table 2-3 by its row total. Thus, Table 2-4 contains estimates of the conditional probability of occurrence of each category of ceiling height 5 hr later when the category at the input time, the season of the year, and the input hour are given. Four tables were computed for each of the 48 predictands. These tables have been published [1]. A forecast is made by entering the proper forecast table with the value of the predictand at the input hour and reading off the five estimated probabilities.

2.4.3 Statistical Techniques

Four statistical techniques were selected for the test. A description of these techniques is given in a technical publication [3]. The grouping, Lund, and multiple-discriminant-analysis techniques produce probability forecasts as their basic output. The Lewis technique produces a categorical forecast. In the case of the Lewis technique, forecasts for only four predictands were produced for the evaluation year by the Electronic Computer Branch, 1210th Weather Squadron, Air Weather Service, and were delivered to The Travelers Research Center for evaluation. Forecasts for all other techniques were prepared by The Travelers Research Center. Because of the limited number of predictands treated by the Lewis technique, their evaluation is described separately (see Section 3.3.3).

2.4.4 Subjective Forecasts

2.4.4.1 Categorical Forecasts

The Forecast Evaluation Working Group chose a cross section of subjective forecasts currently in use by operating agencies. Terminal forecasts prepared under routine operational conditions during the evaluation year (October 1, 1960, through September 30, 1961) were collected and decoded at The Travelers Research Center for all stations except Atlantic City. The Atlantic City forecasts were furnished in decoded form by the U.S. Weather Bureau. The types of forecasts used at each station are given in Table 2-5.

TABLE 2-5
TYPES OF SUBJECTIVE FORECAST
USED IN EVALUATION

Sta	Type of forecast
ACY	FT1
CEF	TAFOR
DCA	FT2, SPF*
IDL	FT1, SPF*
OFF	TAFOR
RND	PLATFS
WRI	TAFOR, SAGE

*Special probability forecast.

FT1s, FT2s, and TAFORs are the usual aviation terminal forecasts prepared by civilian and military personnel. SAGE forecasts are special 2-hr terminal forecasts prepared, at present, 12 times daily by military forecasters for the Air Defense Command. PLATFSs are aviation forecasts prepared at Kansas City, Missouri, for 27 Midwestern Air Force bases.

Subjective categorical forecasts were decoded for only 42 of the 48 predictands.

2.4.4.2 Probability Forecasts

Weather Bureau forecasters at Idlewild and Washington National Airports prepared 3-, 5-, and 7-hr probability forecasts on an experimental basis as part of this test. The forecasters were instructed to use data up to and including the input time and to prepare forecasts not later than 1 hr after this time.

2.4.4.3 Decoding Procedures

Most forecasts were received in the format used in the communication networks and had to be completely decoded. The general principle applied in decoding was to arrive at what the forecaster had in mind at the instants of time selected as the valid times. The development and changes in weather reflected in the forecasts were assumed to be linear. A complete description of the decoding procedure may be found in the supplement to this report.

2.4.5 Input and Valid Times

The input times and valid times to be used when preparing forecasts on the independent data were selected by the Forecast Evaluation Working Group and are given in Table 2-6. SAGE forecasts are issued at the same time as the observation from which the forecast is prepared.

Forecasts prepared by all statistical techniques used the same input and valid times as the subjective forecasts with which they were compared.

2.4.6 Method of Verifying Forecasts

The Forecast Evaluation Working Group specified the method for verifying the forecasts. The plan [5] approved by the FAA stated:

> Forecasts prepared for the PTET [Plan for Test and Evaluation of Terminal Forecasting Techniques] will be expressed in categorical and/or probabilistic terms. For the categorical forecasts, a score devised by Bryan (1961) will be obtained from each contingency table of forecast vs. observed values and will be used to compare forecast techniques. Various other scores such as percent correct, skill scores of various kinds, prefigurance and post-agreement will also be computed. For the probabilistic

TABLE 2-6
INPUT AND VALID TIMES FOR SUBJECTIVE FORECASTS

Sta	Forecast		Input time, EST	Valid time, EST
	Type	Length, hr		
ACY	FT1	3, 5, 7	05	08, 10, 12
			17	20, 22, 00
CEF	TAFOR	2, 4, 6	05	07, 09, 11
			17	19, 21, 23
DCA	FT1 and SPF*	3, 5, 7	05	08, 10, 12
			17	20, 22, 00
IDL	FT2 and SPF*	3, 5, 7	05	08, 10, 12
			11	14, 16, 18
			17	20, 22, 00
			23	02, 04, 06
OFF	TAFOR	2, 4, 6	05	07, 09, 11
			17	19, 21, 23
RND	PLATFS	2, 4, 6	07	09, 11, 13
			19	21, 23, 01
WRI	SAGE	2	05	07
			17	19
	TAFOR	2, 4, 6	05	07, 09, 11
			17	19, 21, 23

*Special probability forecast.

forecasts, the P-score, described by Brier and Allen (1951), will be calculated and compared using the t-test for the means of paired observations.

This plan was followed. Brier-Allen P-scores and Bryan scores for each station, and averages over the seven stations, are presented in this report.

A variety of scores for categorical forecasts other than that devised by Bryan was considered by the Working Group. The Bryan score was designated as the primary scoring procedure for categorical forecasts. The Working Group requested that other scores be computed in the process of evaluation. For the purpose of measuring these scores, all probability forecasts will be converted to categorical forecasts by applying a loss function designed to maximize the scores. These scores will be presented in the supplement.

2.4.6.1 Brier-Allen P-score

For a single probability forecast, the P-score is defined as

$$P = \sum_{i=1}^{5} (f_i - E_i)^2, \tag{2-1}$$

where E_i takes the value 1 or 0 according to whether the predictand occurs in class i or not; f_1, f_2, f_3, f_4, and f_5 represent the forecast probabilities. A P-score of 0 indicates a perfect forecast; the poorest score is 2, which occurs when a probability of 1 is assigned to other than the correct forecast. In comparing two forecasts of the same observed quantity, the lower P-score indicates the better forecast. The P-score is widely used and accepted for verifying probability forecasts.

2.4.6.2 Paired-comparison t-test

Extreme care was taken to ensure that exactly the same forecasts were made in any comparison of forecast techniques. That is, if a 5-hr subjective ceiling forecast was made at 05 Eastern Standard Time (EST) October 12, 1960, then statistical forecasts were made also at this time. There were no exceptions. This matching of forecasts must be done to ensure a valid comparison of forecast techniques. It also permits use of the t-test for paired comparisons. This tests whether the mean P-score for one forecast technique is significantly better than the mean P-score of another technique.

The test value is

$$t = \bar{d} \left[\frac{\sum_{j=1}^{N} (d_j - \bar{d})^2}{N(N-1)} \right]^{-1/2}, \tag{2-2}$$

where $\bar{d}$ is the difference between the mean P-scores, d_j is the difference between individual P-scores, and N is the number of forecasts made. The P-scores of two forecast techniques tend to be highly correlated because poor scores tend to be made by both techniques

on difficult forecast situations and good forecasts tend to be made on easy situations. The power of this widely used test lies in eliminating the correlation by taking differences, and this can be done only when the forecasts are paired.

2.4.6.3 Bryan Score

For a single categorical forecast, the Bryan score is

$$B = W_{ij}, \qquad (2\text{-}3)$$

where W_{ij} is the merit or demerit ascribed to a forecast of category i when category j occurs. The W-values are designed to distinguish between levels of skill in forecasting the occurrence or nonoccurrence of category 5 and, also, in forecasting which non-5 category will occur. In developing the score, it was taken for granted that forecasting methods would be compared by the t-test. Because the Bryan score in its corrected form has not been published previously, a nonmathematical description is given in Appendix A and a mathematical description will be given in a separate publication.*

2.5 Generation of Categorical Forecasts from Probability Forecasts

The statistical forecast techniques, except for the Lewis technique, produce only probability forecasts. To compare these forecasts with persistence, Lewis technique, and subjective forecasts (which are in categorical form only), it is necessary to use the probability forecasts to generate categorical forecasts. If the forecasts were perfect, there would be no question of the method of generation. The correct category would be given by a probability of 1 every time, and that would be the categorical forecast. In the presence of uncertainty in the forecasts, the decision is not this simple.

The Bryan score specifies a loss function and provides a means for arriving at forecasts on the basis of maximizing gain or, in this case, maximizing the score. One of the important features of probability forecasts is that they lend themselves to such treatment. Therefore, the categorical forecasts for all statistical techniques are generated in such a fashion as to maximize the Bryan score. When other scores are computed, they also will be maximized.

This is done as follows. Let f_1, f_2, f_3, f_4, and f_5 be the forecast probabilities. Five quantities are computed:

$$G_1 = \sum_{j=1}^{5} W_{1j} f_j, \quad G_2 = \sum_{j=1}^{5} W_{2j} f_j, \quad \ldots, \quad G_5 = \sum_{j=1}^{5} W_{5j} f_j. \qquad (2\text{-}4)$$

The maximum G gives the categorical forecast. Thus, if G_1 is largest, category 1 is forecast; etc. The W-values are the same as those in Eq. (2-3).

*Bryan, J. G., Scoring System for Categorical Forecasts of Ceiling and Visibility. TRC Tech. Rpt. 7044-59 (FAA Tech. Publication 26) (to be published).

2.6 Summary of the Test Plan

The test was designed to compare the forecasting accuracy of the statistical techniques with one another and with subjective forecasts prepared under current operating conditions. A secondary objective was to evaluate the accuracy of subjective probability forecasts, which are not now prepared in routine operation. The control technique for categorical forecasts was persistence and, for probability forecasts, climatological expectancy of persistence.

A dependent sample of 10 years of data was used to develop the statistical techniques. Categorical and probability forecasts were produced with these techniques on one year of independent data. Subjective categorical forecasts for the same year were collected and decoded. Special subjective probability forecasts were made. The forecasting skills of all probability forecasts were compared by means of the P-score; categorical forecasts were compared by means of the Bryan score.

Three statistical techniques produced probability forecasts: grouping, Lund, and multiple-discriminant analysis. These were compared on 48 predictands: ceiling and visibility forecasts at Idlewild and Washington National for 2, 3, 5, and 7 hr; at Westover AFB for 2, 3, 4, and 6 hr; at Atlantic City for 3, 5, and 7 hr; and at McGuire, Offutt, and Randolph AFB for 2, 4, and 6 hr.

Special subjective probability forecasts were prepared for ceiling and visibility at Idlewild and Washington National for 3, 5, and 7 hr. These were compared with forecasts made with the three statistical techniques.

Subjective categorical forecasts were collected for 42 of the 48 predictands. The predictands omitted were ceiling and visibility at Idlewild and Washington National for 12 hr and at Westover for 3 hr. Categorical forecasts for the 42 predictands were generated from the probability forecasts of the three statistical techniques. The categorical forecasts for the six techniques were compared. A separate evaluation of the Lewis technique was necessitated by the small sample of forecasts available.

3.0 SUMMARY OF TEST RESULTS

3.1 Probability Forecasts by Statistical Techniques

Probability forecasts of all 48 predictands were made by three statistical techniques and the control technique (climatological expectancy of persistence, CEP) and were verified by the P-score. These P-scores are given in Table 3-1. Examination of Table 3-1 shows that the multiple-discriminant-analysis (MDA) technique yielded the best score on 44 of the predictands and CEP, the control technique, yielded the best score on the other 4. The paired-comparison t-test probabilities were combined by Fisher's method [4] to obtain a single significance test for all 48 predictands. The MDA P-scores were significantly better than those of CEP beyond the 1% level of significance.

An index (I) of the amount of increase or decrease in forecasting skill of statistical techniques relative to the control technique is

$$I = \frac{P_{CEP} - P_S}{P_{CEP} - 0} \times 100, \qquad (3\text{-}1)$$

where P_{CEP} and P_S indicate CEP and statistical-technique P-scores respectively. Because 0 is a perfect P-score, $P_{CEP} - 0$ is the total amount of forecasting skill not accounted for by CEP. Therefore, I is the ratio of the improvement or deterioration in forecasting skill of a statistical technique relative to the control in terms of forecasting skill remaining to be accounted for. The factor 100 puts I into percentage form.

Values of I were computed for the "average-over-station" P-scores of Table 3-1 and are presented in Table 3-2. The results displayed in Tables 3-1 and 3-2 and the associated t-tests indicate that MDA probability forecasts of 2- to 7-hr ceiling and visibility are significantly better than probability forecasts made either by the other two statistical techniques or by the control technique, as measured by the Brier-Allen P-score.

3.2 Bryan Scores of Categorical Forecasts by All Techniques

The probability forecasts of the four statistical techniques (including the control technique) were used to generate categorical forecasts for 42 of the 48 predictands in such a way as to maximize the Bryan score (see Section 2.5). Categorical persistence and subjective forecasts for the same 42 predictands were also available. The forecasting skill of each technique on every predictand was measured by the Bryan score, and the results are given in Table 3-3. Paired-comparison t-tests were made between the scores attained by each technique and every other technique.

Table 3-3 indicates that the Bryan scores for MDA were highest on 29 of the 42 predictands, those for subjective forecasts were highest on 6 predictands, and those for the other techniques were highest on the remaining 7 predictands. The average Bryan scores for all predictands, given in Table 3-3(c), indicate that MDA scores were highest and CEP and subjective scores were approximately equal and ranked second. The paired-

TABLE 3-1
P-SCORE* TEST RESULTS FOR THREE STATISTICAL TECHNIQUES
FOR EVALUATION-YEAR DATA

(a) The predictand element is ceiling

Predictand		No. of fcsts	P-score			
Sta	Fcst length, hr		CEP	Group	Lund	MDA
ACY	3	682	0.1783	0.2148	0.3164	0.1755
CEF	2	651	0.2178	0.2153	0.2897	0.1947
CEF	3	661	0.2681	0.2572	0.2577	0.2445
DCA	2	657	0.1222	0.1255	0.2199	0.1172
DCA	3	543	0.1349	0.1372	0.2417	0.1210
IDL	2	1,454	0.1816	0.1885	0.2082	0.1732
IDL	3	1,290	0.1825	0.1849	0.3446	0.1701
OFF	2	697	0.1700	0.1769	0.2588	0.1602
RND	2	693	0.2937	0.2949	0.3484	0.2707
WRI	2	726	0.2583	0.2401	0.2628	0.2208
Mean	2-3	8,054	0.2007	0.2035	0.2748	0.1848
ACY	5	652	0.2106	0.2098	0.3582	0.1969
CEF	4	657	0.2936	0.2897	0.3471	0.2465
DCA	5	546	0.1368	0.1360	0.2290	0.1213
IDL	5	1,457	0.2452	0.2385	0.2838	0.2171
OFF	4	656	0.2314	0.2541	0.2854	0.1945
RND	4	681	0.3508	0.3808	0.4261	0.3246
WRI	4	610	0.2907	0.2781	0.3190	0.2396
Mean	4-5	5,259	0.2513	0.2553	0.3212	0.2201
ACY	7	676	0.2614	0.2372	0.3681	0.2284
CEF	6	672	0.3335	0.3405	0.3715	0.2854
DCA	7	682	0.1681	0.1616	0.2026	0.1402
IDL	7	1,459	0.2788	0.2649	0.2857	0.2578
OFF	6	668	0.2448	0.2627	0.2926	0.2017
RND	6	679	0.3872	0.4362	0.4949	0.3559
WRI	6	608	0.3271	0.3501	0.4299	0.2810
Mean	6-7	5,444	0.2858	0.2933	0.3493	0.2501

*Lower score indicates better forecast.

(b) The predictand element is visibility

Predictand		No. of fcsts	P-score			
Sta	Fcst length, hr		CEP	Group	Lund	MDA
ACY	3	700	0.1895	0.2095	0.2477	0.1937
CEF	2	685	0.2443	0.2093	0.3852	0.1768
CEF	3	679	0.2103	0.2308	0.2620	0.2254
DCA	2	687	0.1519	0.1530	0.1514	0.1394
DCA	3	611	0.1237	0.1282	0.1296	0.1208
IDL	2	1,366	0.1424	0.1488	0.1802	0.1553
IDL	3	1,343	0.1530	0.1616	0.1516	0.1425
OFF	2	698	0.1107	0.1102	0.1332	0.1099
RND	2	703	0.1194	0.1259	0.2255	0.1116
WRI	2	728	0.2619	0.2593	0.2722	0.2361
Mean	2-3	8,200	0.1707	0.1737	0.2139	0.1612
ACY	5	664	0.1468	0.1530	0.1605	0.1506
CEF	4	676	0.2454	0.2580	0.3750	0.2306
DCA	5	699	0.1116	0.1197	0.1276	0.1101
IDL	5	1,317	0.1580	0.1639	0.1961	0.1515
OFF	4	698	0.1253	0.1347	0.1298	0.1212
RND	4	683	0.0807	0.0991	0.2521	0.0779
WRI	4	607	0.3048	0.3086	0.3054	0.2707
Mean	4-5	5,344	0.1675	0.1767	0.2209	0.1589
ACY	7	657	0.1474	0.1529	0.2558	0.1366
CEF	6	642	0.2682	0.2837	0.4550	0.2671
DCA	7	714	0.0970	0.1047	0.1048	0.0933
IDL	7	1,345	0.1875	0.1962	0.1893	0.1788
OFF	6	622	0.1124	0.1167	0.1789	0.1099
RND	6	683	0.0907	0.0973	0.2814	0.0843
WRI	6	608	0.2763	0.3070	0.3171	0.2678
Mean	6-7	5,271	0.1685	0.1798	0.2546	0.1625

(c) Composite of (a) and (b) for all stations and forecasts

No. of fcsts	Mean P-score			
	CEP	Group	Lund	MDA
37,572	0.2047	0.2106	0.2689	0.1875

TABLE 3-2
PERCENT IMPROVEMENT (I) OF P-SCORES RELATIVE TO CEP
FOR EVALUATION-YEAR DATA

Predictand		No. of fcsts	I, %		
Elem	Fcst length, hr		Group	Lund	MDA
CIG	2-3	8054	-1.4	-36.9	7.9
CIG	4-5	5259	-1.6	-27.8	12.4
CIG	6-7	5444	-2.6	-22.2	12.5
VIS	2-3	8200	-1.8	-25.3	5.6
VIS	4-5	5344	-5.5	-31.9	5.1
VIS	6-7	5271	-6.7	-51.1	3.6

comparison t-test showed that the MDA scores were statistically significantly higher than CEP and subjective scores. At the 5% level, MDA scores were higher than CEP scores on 18 of the 42 predictands and higher than the subjective forecast scores on 13 predictands. Fisher's test [4] indicated that, for all forecasts combined, the scores achieved by the MDA technique were statistically significantly better than those of either CEP or subjective techniques, beyond the 1% level.

The average Bryan score for a very large number of perfect forecasts would be 1.0. However, for any limited number of forecasts, the maximum attainable score must be computed from the observed frequencies of the various categories. The index I defined in Eq. (3-1) now becomes

$$I = \frac{B_S - B_C}{B_M - B_C} \times 100, \qquad (3\text{-}2)$$

where B_S is the average Bryan score achieved by a forecast technique, B_C is the average Bryan score achieved by the control technique, and B_M is the maximum average Bryan score attainable within the sample of forecasts. Values of I were computed for the "average-over-station" Bryan scores taken from Table 3-3 and are presented in Table 3-4. MDA yielded the highest value of the index I.

TABLE 3-3
BRYAN-SCORE* TEST RESULTS FOR CATEGORICAL FORECASTS
FOR EVALUATION-YEAR DATA

(a) Predictand element is ceiling

Predictand		No. of fcsts	Bryan score					
Sta	Fcst length, hr		Pers	Subj	CEP	Group	Lund	MDA
ACY	3	681	0.317	0.343	0.310	0.203	0.347	0.364
CEF	2	650	0.464	0.462	0.519	0.496	0.418	0.505
DCA	3	541	0.341	0.337	0.270	0.344	0.375	0.391
IDL	3	1,283	0.415	0.426	0.424	0.339	0.385	0.446
OFF	2	694	0.597	0.540	0.597	0.596	0.572	0.613
RND	2	692	0.608	0.585	0.620	0.577	0.552	0.620
WRI	2	611	0.440	0.400	0.421	0.481	0.399	0.535
Mean	2-3	5,152	0.455	0.442	0.452	0.434	0.435	0.496
ACY	5	651	0.231	0.299	0.232	0.283	0.190	0.330
CEF	4	655	0.386	0.374	0.363	0.362	0.336	0.425
DCA	5	544	0.229	0.261	0.267	0.301	0.288	0.335
IDL	5	1,451	0.374	0.383	0.375	0.395	0.462	0.454
OFF	4	652	0.360	0.378	0.345	0.003	0.369	0.426
RND	4	680	0.361	0.447	0.406	0.359	0.398	0.432
WRI	4	519	0.302	0.317	0.321	0.326	0.326	0.383
Mean	4-5	5,152	0.320	0.351	0.330	0.290	0.338	0.398
ACY	7	676	0.194	0.271	0.184	0.258	0.210	0.246
CEF	6	669	0.283	0.400	0.325	0.321	0.343	0.411
DCA	7	681	0.164	0.250	0.204	0.311	0.315	0.306
IDL	7	1,452	0.254	0.360	0.264	0.336	0.323	0.342
OFF	6	664	0.323	0.328	0.298	0.320	0.329	0.372
RND	6	678	0.244	0.408	0.268	0.219	0.259	0.335
WRI	6	519	0.228	0.347	0.264	0.247	0.187	0.356
Mean	6-7	5,339	0.241	0.338	0.258	0.287	0.281	0.338

*Higher score indicates better forecast.

(b) Predictand element is visibility

Predictand		No. of fcsts	Bryan score					
Sta	Fcst length, hr		Pers	Subj	CEP	Group	Lund	MDA
ACY	3	699	0.242	0.254	0.208	0.288	0.285	0.281
CEF	2	682	0.404	0.489	0.357	0.539	0.067	0.644
DCA	3	609	0.186	0.216	0.495	0.288	0.183	0.498
IDL	3	1,336	0.298	0.426	0.393	0.376	0.364	0.535
OFF	2	694	0.356	0.300	0.284	0.294	0.319	0.382
RND	2	702	0.331	0.310	0.296	0.423	0.492	0.547
WRI	2	612	0.529	0.513	0.543	0.512	0.460	0.588
Mean	2-3	5,334	0.335	0.358	0.368	0.389	0.310	0.496
ACY	5	663	0.094	0.221	0.095	0.077	0.108	0.166
CEF	4	672	0.324	0.331	0.304	0.343	0.386	0.430
DCA	5	696	0.185	0.087	0.376	0.279	0.305	0.472
IDL	5	1,311	0.180	0.248	0.338	0.270	0.180	0.359
OFF	4	693	0.295	0.200	0.277	0.332	0.396	0.364
RND	4	682	0.148	0.217	0.220	0.154	0.205	0.310
WRI	4	514	0.283	0.299	0.323	0.290	0.377	0.410
Mean	4-5	5,231	0.216	0.229	0.276	0.249	0.280	0.359
ACY	7	657	0.039	0.166	0.064	0.033	0.051	0.052
CEF	6	639	0.223	0.222	0.246	0.372	0.298	0.307
DCA	7	713	0.110	0.062	0.143	0.272	-0.008	0.274
IDL	7	1,338	0.129	0.233	0.278	0.049	0.275	0.347
OFF	6	618	0.184	0.085	0.191	0.271	0.267	0.307
RND	6	682	0.035	0.108	0.174	0.093	0.096	0.193
WRI	6	517	0.242	0.198	0.262	0.234	0.254	0.248
Mean	6-7	5,164	0.137	0.153	0.194	0.189	0.176	0.247

(c) Composite of (a) and (b) for all stations and forecasts

No. of fcsts	Mean Bryan score					
	Pers	Subj	CEP	Group	Lund	MDA
31,372	0.284	0.312	0.313	0.306	0.303	0.389

TABLE 3-4

PERCENT IMPROVEMENT (I) OF BRYAN SCORES FOR CATEGORICAL FORECASTS RELATIVE TO PERSISTENCE FOR EVALUATION-YEAR DATA

Predictand		I, %				
Elem	Fcst length, hr	Subj	CEP	Group	Lund	MDA
CIG	2-3	- 3.2	-0.7	-5.1	-4.8	9.9
CIG	4-5	5.6	1.8	-5.4	3.3	14.2
CIG	6-7	14.7	2.6	7.0	6.1	14.7
VIS	2-3	2.7	3.9	6.4	-3.0	19.2
VIS	4-5	1.7	7.7	4.2	8.2	18.2
VIS	6-7	1.9	6.9	6.3	4.7	13.3

3.3 Special Evaluations

A number of special or experimental evaluations were either requested by the Working Group or were required because of peculiarities encountered in the course of the evaluation. Among these were the SAGE forecast evaluation requested by the Air Force, the evaluation of special subjective probability forecasts made by the U. S. Weather Bureau, and the evaluation of terminal forecasts made by the Lewis technique.

3.3.1 SAGE Forecasts

Subjective SAGE 2-hr ceiling and visibility forecasts at McGuire AFB were collected and decoded. These were categorical forecasts and, therefore, were compared with categorical forecasts produced by all other techniques, including the subjective TAFOR made for the same station. The Bryan scores are displayed in Table 3-5.

3.3.2 Subjective Probability Forecasts

Special subjective probability forecasts were prepared at Washington National and Idlewild Airports by U.S. Weather Bureau forecasters. The predictands for both stations were 3-, 5-, and 7-hr visibility and ceiling. The instructions given to the forecasters were that the forecasts were to be made not later than 1 hr after the input time. These probability forecasts were verified and compared with other techniques by means of the P-score.

TABLE 3-5
BRYAN SCORES FOR THE EVALUATION OF 594 TWO-hr SAGE TERMINAL FORECASTS FOR McGUIRE AFB

Elem	Bryan score						
	SAGE	Pers	Subj	CEP	Group	Lund	MDA
CIG	0.408	0.451	0.411	0.432	0.483	0.411	0.545
VIS	0.347	0.542	0.526	0.556	0.524	0.476	0.600

In addition, categorical forecasts were made from these probability forecasts in such a manner as to maximize the Bryan score. These were verified by means of the Bryan score.

3.3.2.1 Brier-Allen P-score Verification

The Brier-Allen P-scores for the 12 predictands are given in Table 3-6 for the subjective probability forecasts and for corresponding forecasts made by four statistical techniques. The P-scores for subjective ceiling forecasts are lower (better) than any of the statistical techniques except MDA. This is an interesting result when it is realized that the subjective forecasters did not have much experience in preparing probability forecasts. The paired-comparison t-test showed that, for all ceiling forecasts combined, MDA P-scores are statistically significantly better than those of the subjective forecasts beyond the 1% level. The P-scores for the subjective probability forecasts of visibility were not as good as those of the statistical techniques.

3.3.2.2 Bryan-score Verification

The subjective probability forecasts were converted to categorical forecasts by maximizing the Bryan score. The Bryan scores for the 12 predictands for five techniques are given in Table 3-7. These subjective categorical forecasts yielded higher scores than any other technique, for both ceiling and visibility. The paired-comparison t-test indicated that the improvement in Bryan scores of the subjective forecasts over those of the second-ranking technique (MDA) is statistically significant beyond the 2% level.

These results differ from the results obtained with the P-score verification. This may be because the Bryan score puts great emphasis upon correct forecasts in the non-5 categories, whereas the P-score treats the categories equally. Thus, probability forecasts that are quite good at specifying probabilities in the non-5 categories but not so good in category 5 will receive a poorer P-score but a better Bryan score.

TABLE 3-6
EVALUATION OF SUBJECTIVE PROBABILITY FORECASTS:
BRIER-ALLEN P-SCORES* FOR 12 PREDICTANDS FOR EVALUATION-YEAR DATA

(a) Predictand element is ceiling

Predictand		No. of fcsts	Brier-Allen P-score				
Sta	Fcst length, hr		Subj	CEP	Group	Lund	MDA
DCA	3	539	0.1371	0.1359	0.1382	0.2435	0.1219
DCA	5	540	0.1436	0.1375	0.1373	0.2273	0.1225
DCA	7	677	0.1688	0.1688	0.1627	0.2041	0.1412
IDL	3	1,189	0.1648	0.1796	0.1808	0.3408	0.1651
IDL	5	1,332	0.2282	0.2411	0.2330	0.2763	0.2111
IDL	7	1,333	0.2547	0.2740	0.2589	0.2781	0.2507
Mean	3-7	5,610	0.1829	0.1895	0.1852	0.2617	0.1688

(b) Predictand element is visibility

Predictand		No. of fcsts	Brier-Allen P-score				
Sta	Fcst length, hr		Subj	CEP	Group	Lund	MDA
DCA	3	605	0.1347	0.1250	0.1294	0.1309	0.1220
DCA	5	694	0.1330	0.1124	0.1206	0.1284	0.1109
DCA	7	708	0.1237	0.0951	0.1029	0.1029	0.0914
IDL	3	1,230	0.1487	0.1409	0.1492	0.1401	0.1297
IDL	5	1,208	0.1732	0.1513	0.1579	0.1876	0.1457
IDL	7	1,231	0.1825	0.1729	0.1816	0.1754	0.1650
Mean	3-7	5,676	0.1493	0.1329	0.1403	0.1442	0.1275

*Lower score indicates better forecast.

TABLE 3-7
EVALUATION OF EXPERIMENTAL SUBJECTIVE CATEGORICAL FORECASTS:
BRYAN SCORES* FOR 12 PREDICTANDS FOR EVALUATION-YEAR DATA

(a) Predictand element is ceiling

Predictand		No. of fcsts	Bryan score				
Sta	Fcst length, hr		Pers	Subj	Special subj	CEP	MDA
DCA	3	537	0.343	0.339	0.399	0.272	0.393
DCA	5	538	0.232	0.264	0.351	0.270	0.338
DCA	7	676	0.165	0.252	0.370	0.205	0.308
IDL	3	1,182	0.428	0.438	0.477	0.432	0.459
IDL	5	1,326	0.387	0.391	0.503	0.385	0.460
IDL	7	1,327	0.262	0.353	0.464	0.273	0.360
Mean	3-7	5,586	0.303	0.340	0.427	0.306	0.386

(b) Predictand element is visibility

Predictand		No. of fcsts	Bryan score				
Sta	Fcst length, hr		Pers	Subj	Special subj	CEP	MDA
DCA	3	603	0.188	0.218	0.586	0.500	0.502
DCA	5	691	0.186	0.088	0.409	0.378	0.475
DCA	7	707	0.112	0.063	0.260	0.144	0.277
IDL	3	1,223	0.285	0.385	0.501	0.324	0.480
IDL	5	1,202	0.168	0.243	0.381	0.340	0.350
IDL	7	1,225	0.126	0.216	0.411	0.238	0.318
Mean	3-7	5,651	0.178	0.202	0.425	0.321	0.400

*Higher score indicates better forecast.

TABLE 3-8
COMPARISON OF TWO TYPES OF SUBJECTIVE CATEGORICAL FORECAST:
COMPOSITE CONTINGENCY TABLES FOR 12 PREDICTANDS

(a) Categorical forecasts made by subjective forecasters under operating conditions

		Observed					Total
		1	2	3	4	5	
Forecast	1	12	7	2	2	7	30
	2	51	78	41	20	69	259
	3	29	67	140	107	143	486
	4	11	31	89	203	317	651
	5	44	39	91	263	9,374	9,811
Total		147	222	363	595	9,910	11,237

(b) Categorical forecasts generated from special subjective probability forecasts to maximize number of hits

		Observed					Total
		1	2	3	4	5	
Forecast	1	32	18	12	3	20	85
	2	44	79	42	24	71	260
	3	22	60	148	108	112	450
	4	6	24	82	218	299	629
	5	43	41	79	242	9,408	9,813
Total		147	222	363	595	9,910	11,237

3.3.2.3 Comparison of Two Types of Subjective Forecast

It is clear from Table 3-7 that the Bryan scores for subjective forecasts generated from forecast probabilities are better than those of the subjective categorical forecasts prepared routinely. To gain further information, additional categorical forecasts were produced from the subjective forecast probabilities. This was done by specifying that the categorical forecast is that given by the category with the highest probability. That is, if f_1, f_2, f_3, f_4, and f_5 are forecast probabilities, then the category with the highest f is the categorical forecast. This method of generating categorical forecasts is designed to maximize the number of hits. Table 3-8 compares a composite contingency table over all 12 predictands for this type of subjective categorical forecast with a similar table for routinely prepared subjective categorical forecasts. The number of hits is 9807 for the routine subjective forecasts, and 9385 for the categorical forecasts generated from the probabilities.

3.3.3 Lewis-technique Forecasts

Forecasts on the evaluation-year data were made by the Electronic Computer Branch, 1210th Weather Squadron, Air Weather Service, and delivered to The Travelers Research Center, Inc., for verification. The only forecasts included were those made once a day, at 0500 EST, for 5- and 7-hr ceiling and 2- and 5-hr visibility at Washington National Airport. The Lewis technique produces only categorical forecasts. Therefore, the forecasts were verified by means of the Bryan score. The results are presented in Table 3-9.

The Bryan scores for ceiling forecasts produced by the Lewis technique are higher than those produced by persistence, subjective, CEP, or Lund, about the same as those produced by MDA, and lower than those produced by grouping. The paired-comparison t-test indicates that none of the scores for ceiling forecasts is significantly different from

TABLE 3-9
EVALUATION BY BRYAN SCORES OF LEWIS TERMINAL FORECASTS
FOR WASHINGTON NATIONAL AIRPORT

Predictand		No. of fcsts	Bryan scores						
Elem	Fcst length, hr		Lewis	Pers	Subj	CEP	Group	Lund	MDA
CIG	5	225	0.416	0.262	0.294	0.325	0.415	0.321	0.373
CIG	7	341	0.275	0.224	0.223	0.303	0.454	0.254	0.327
Mean	2-7	566	0.346	0.243	0.259	0.314	0.435	0.288	0.350
VIS	2	344	0.459	0.509	*	0.462	0.821	0.498	1.327
VIS	5	342	0.274	0.223	0.137	0.658	0.393	0.435	0.667
Mean	2-5	686	0.367	0.366	—	0.560	0.607	0.467	0.997

*No forecasts avilable.

any other score at the 5% level. This is an indication that the sample is too small to allow any reasonable judgment. Lewis visibility scores were lower than those produced by the other techniques, with the exception of persistence and subjective forecasts. A significant difference at the 1% level was found only between the Lewis scores and the MDA scores.

4.0 REFERENCES

1. Ceiling and Visibility Contingency Tables for Persistence and Climatological Expectancy of Persistence for Selected Stations, 112 pp. The Travelers Research Center, Inc., Hartford, 1962.

2. Common Aviation Weather System Design, 181 pp. Rpt. 1, Federal Aviation Agency, Washington, D.C., 1962.

3. Enger, I., Statistical Forecasting Under Operational Conditions, 29 pp. Tech. Publ. 12, Contract FAA/BRD-363, The Travelers Research Center, Inc., Hartford, Jun. 1962.

4. Fisher, R. A., Statistical Methods for Research Workers, 11th ed., pp. 99–101. New York: Hafner Publishing Co., 1950.

5. Reed, L. J., A. H. Murphy, and I. Enger, A Plan for the Test and Evaluation of Terminal Forecasting Techniques (PTET), 40 pp. Tech. Memo. 1, Contract FAA/BRD-363, The Travelers Research Center, Inc., Hartford, 1961.

APPENDIX A. DESCRIPTION OF THE BRYAN SCORE*

Both ceiling and visibility have been subdivided into five operationally significant classes. Each predictand, therefore, can yield just 25 possible combinations of observed and forecast classes. The problem of scoring the categorical forecasts is one of deciding upon a quantitative merit or demerit for each combination. Concensus has proved so difficult to achieve, however, that it has been agreed to compromise on a scoring system developed mathematically from reasonable but arbitrary assumptions. The assumptions themselves have been arrived at through a gradual process of exploration and amendment.

With probability of occurrence heavily concentrated in one class, it was found that the prescribed conditions in their original form could not be satisfied simultaneously, but that an amended set of conditions should be satisfied simultaneously. The original set of conditions will be described first, and, afterward, the amended set.

1. The first assumption was that demerits should be progressive. If the error of forecasting class 2 when in fact class 1 is observed receives the demerit $-d_1$, and the error of forecasting class 3 when class 2 is observed receives the demerit $-d_2$, then the error of forecasting class 3 when class 1 is observed (being regarded as the sum of the two errors) receives the demerit $-(d_1 + d_2)$. It is not necessary under this condition (but it is, as a consequence of other conditions) to assume symmetry. That is, conceivably the error of forecasting class 1 when class 2 actually occurs need not receive the same demerit as the error of forecasting class 2 when class 1 actually occurs. The general pattern of merits (x_1, x_2, ...) and demerits ($-d_1$, $-\delta_1$, $-d_2$, $-\delta_2$, ...) is displayed in Table A-1. If the x's are to have the effect of merits and the -d's and -δ's are to have the effect of demerits, no x, d, or δ can be negative.

2. The second assumption was that if forecasts produced by a given procedure are distributed independently of the observed weather, the forecasts should receive, in the long run, an average score of zero. This assumption was based on the fact that such

*Joseph G. Bryan.

TABLE A-1
ASYMMETRICAL PATTERN OF MERITS AND DEMERITS

Class		Forecast 1	Forecast 2	Forecast 3	Forecast 4	Forecast 5
Observed	1	x_1	$-d_1$	$-(d_1 + d_2)$	$-(d_1 + d_2 + d_3)$	$-(d_1 + d_2 + d_3 + d_4)$
	2	$-\delta_1$	x_2	$-d_2$	$-(d_2 + d_3)$	$-(d_2 + d_3 + d_4)$
	3	$-(\delta_1 + \delta_2)$	$-\delta_2$	x_3	$-d_3$	$-(d_3 + d_4)$
	4	$-(\delta_1 + \delta_2 + \delta_3)$	$-(\delta_2 + \delta_3)$	$-\delta_3$	x_4	$-d_4$
	5	$-(\delta_1 + \delta_2 + \delta_3 + \delta_4)$	$-(\delta_2 + \delta_3 + \delta_4)$	$-(\delta_3 + \delta_4)$	$-\delta_4$	x_5

forecasts are technically devoid of information, in the sense of reducing uncertainty.

3. The third assumption was a corollary of the second. A purely random, but climatologically realistic forecast, being statistically independent of the actual weather, cannot yield uncertainty-reducing information on any observed category. Hence the third assumption was that if a population of such forecasts were subclassified according to the respective categories of the observed weather, the average score of those random forecasts should be zero in <u>each</u> category of the observed.

4. The fourth assumption was based on consideration of a statistical test by which two forecasting methods could be compared. With the purpose of making the statistical test of comparative merit as sensitive as possible, an optimization criterion was imposed. In general, the most sensitive test would have to utilize knowledge of the joint probabilities with which any two methods yield any possible combination of forecasts for a given observed category. Such knowledge is unavailable, but in its place a particular pair of forecast types, representing the extremes of merit, were chosen as the basis of optimization. These were the perfect forecast and the random climatological forecast. The fourth assumption, then, was that the merits and demerits, subject to other conditions, should be determined so as to maximize the value of the statistic defined by the t-test for comparative merit of the perfect and random climatological forecasts. Obviously, if these two forecasts could be identified respectively as perfect and random, in advance, no test would

TABLE A-2
SYMMETRICAL PATTERN OF MERITS AND DEMERITS

Class		Forecast 1	2	3	4	5
Observed	1	x_1	$-d_1$	$-(d_1 + d_2)$	$-(d_1 + d_2 + d_3)$	$-(d_1 + d_2 + d_3 + d_4)$
	2	$-d_1$	x_2	$-d_2$	$-(d_2 + d_3)$	$-(d_2 + d_3 + d_4)$
	3	$-(d_1 + d_2)$	$-d_2$	x_3	$-d_3$	$-(d_3 + d_4)$
	4	$-(d_1 + d_2 + d_3)$	$-(d_2 + d_3)$	$-d_3$	x_4	$-d_4$
	5	$-(d_1 + d_2 + d_3 + d_4)$	$-(d_2 + d_3 + d_4)$	$-(d_3 + d_4)$	$-d_4$	x_5

be necessary; but they are used only as a basis of optimizing scores on which other kinds of forecast may be judged.

5. The fifth assumption was that the perfect forecast should receive, in the long run, an average score of unity. Any other arbitrary positive constant, such as 100%, would satisfy mathematical necessities just as well; but some arbitrary constant has to be chosen, to establish the scale of measurement. Unity was chosen.

The combined effect of assumptions 2 and 3 was to produce a symmetrical pattern of demerits; that is, the mathematical consequence was to make $\delta_1 = d_1$, $\delta_2 = d_2$, $\delta_3 = d_3$, and $\delta_4 = d_4$. The resulting scheme is shown in Table A-2.

The foregoing five assumptions constitute a self-consistent system of conditions, provided that the probabilities of the separate weather classes are not too disparate. With nearly every predictand in our data, however, they are too disparate. Excepting only visibility at Atlantic City, the logical requirement that the d's be nonnegative, in order that the -d's have the effect of demerits, forced all but d_4 to take the value zero. Thus the scoring scheme for every predictand but visibility at Atlantic City reduced to the form illustrated in Table A-3. As to the exception, there are two nonvanishing terms, d_1 and d_4; but the magnitude of d_1 is too small to have an appreciable effect on the scores. The only distinction is that with the other predictands, d_1 is precisely zero, whereas here d_1 is nearly zero.

With d_1, d_2, and d_3 reducing to zero, the merits x_1, x_2, ..., x_5 are determined under assumption 2 as fixed multiples of d_4, thus precluding any attention to assumption 4. This

TABLE A-3
EVANESCED PATTERN OF MERITS AND DEMERITS
FORCED BY VANISHING OF d_1, d_2, and d_3

Class		Forecast 1	2	3	4	5
Observed	1	X_1	0	0	0	$-d_4$
	2	0	X_2	0	0	$-d_4$
	3	0	0	X_3	0	$-d_4$
	4	0	0	0	X_4	$-d_4$
	5	$-d_4$	$-d_4$	$-d_4$	$-d_4$	X_5

exclusion of assumption 4 can be avoided by revising assumptions 1, 2, and 3. When this is done, all assumptions of the amended system are compatible.

Before stating the revisions of assumptions 1, 2, and 3, the exploration of another avenue, via assumption 1, will be described. The pattern of Table A-3 suggests the possibility that the build-up of demerits under assumption 1 might have something to do with narrowing the only permissible solution to $d_1 = 0$, $d_2 = 0$, and $d_3 = 0$. Another scoring pattern that would allow some distinctions to be made, without differentiating among as many degrees of error as Table A-2, is displayed in Table A-4. Unfortunately, when the other assumptions were kept the same, the outcome was identical: $d_1 = 0$, $d_2 = 0$, and $d_3 = 0$.

By this time, it seemed appropriate to start with the pattern of Table A-3 as the revised assumption 1, and seek a modification of assumptions 2 and 3 that would preserve their original intent as far as possible and yet leave room for statistical optimization. A clue may be gleaned from the probability distributions, shown in Table A-5, where p_1, p_2, ..., p_5 are the respective empirical probabilities of the five weather classes at each station. In view of the great predominance of class 5, it was concluded that the intent of assumptions 2 and 3 could be served, to a practical approximation, by confining the appli-

TABLE A-4
NESTED PATTERN OF MERITS AND DEMERITS

Class		Forecast				
		1	2	3	4	5
Observed	1	X_1	$-d_1$	$-d_2$	$-d_3$	$-d_4$
	2	$-d_1$	X_2	$-d_2$	$-d_3$	$-d_4$
	3	$-d_2$	$-d_2$	X_3	$-d_3$	$-d_4$
	4	$-d_3$	$-d_3$	$-d_3$	X_4	$-d_4$
	5	$-d_4$	$-d_4$	$-d_4$	$-d_4$	X_5

TABLE A-5
EMPIRICAL PROBABILITIES

Sta	Elem	p_1	p_2	p_3	p_4	p_5
ACY	CIG	0.028	0.040	0.057	0.083	0.792
ACY	VIS	0.031	0.019	0.034	0.037	0.879
CEF	CIG	0.017	0.029	0.083	0.186	0.685
CEF	VIS	0.020	0.021	0.069	0.080	0.810
DCA	CIG	0.004	0.015	0.048	0.076	0.857
DCA	VIS	0.005	0.005	0.016	0.032	0.942
IDL	CIG	0.010	0.029	0.052	0.087	0.822
IDL	VIS	0.010	0.013	0.021	0.033	0.923
OFF	CIG	0.008	0.047	0.034	0.134	0.777
OFF	VIS	0.007	0.009	0.027	0.033	0.924
RND	CIG	0.016	0.022	0.132	0.147	0.683
RND	VIS	0.010	0.008	0.025	0.025	0.932
WRI	CIG	0.020	0.038	0.084	0.139	0.719
WRI	VIS	0.022	0.024	0.089	0.096	0.769

cation of original assumption 2 to class 5 and relaxing assumption 3 to the extent of giving a random climatological forecast an over-all average score of zero rather than by requiring the average score to be zero separately in each observed class. Taken together, these assumptions make the average score of a random climatological forecast zero in observed class 5 and also zero in the other four classes combined. There are six constants and three equations of constraint; hence there are three degrees of freedom left for optimization. An informationless forecast produced by the best theory-of-games pure strategy can earn a nonzero average score, but the maximum average is small. In this connection, it is well to remember that the t-test eliminates the effect of a common base in paired comparisons, and that probability forecasts actually take advantage of the game-theory principle but apply it to presumably sharper probabilities.

The revised assumptions are as follows.

1. The scheme of merits and demerits are as displayed in Table A-6.

2. Any procedure that produces forecasts of class 5 independently of the occurrence of class 5 receives a population average score of zero on observed class 5.

3. Taken over all classes, a random climatological forecast receives a population average score of zero.

TABLE A-6
FINALIZED PATTERN OF MERITS AND DEMERITS

Class		Forecast 1	2	3	4	5
Observed	1	X_1	0	0	0	-Y
	2	0	X_2	0	0	-Y
	3	0	0	X_3	0	-Y
	4	0	0	0	X_4	-Y
	5	-Y	-Y	-Y	-Y	X_5

TABLE A-7
NUMERICAL VALUES OF MERITS AND DEMERITS

Sta	Elem	X_1	X_2	X_3	X_4	X_5	Y
ACY	CIG	4.296	4.350	4.432	4.569	0.090	0.342
ACY	VIS	7.991	7.887	8.020	8.050	0.040	0.290
CEF	CIG	2.357	2.381	2.500	2.766	0.246	0.535
CEF	VIS	4.642	4.648	4.896	4.958	0.090	0.385
DCA	CIG	6.188	6.262	6.478	6.671	0.074	0.446
DCA	VIS	16.335	16.388	16.565	16.840	0.025	0.404
IDL	CIG	4.922	5.021	5.146	5.346	0.087	0.404
IDL	VIS	12.462	12.497	12.601	12.769	0.027	0.323
OFF	CIG	3.635	3.778	3.728	4.141	0.145	0.506
OFF	VIS	12.489	12.520	12.746	12.837	0.029	0.355
RND	CIG	2.336	2.399	2.667	2.711	0.230	0.493
RND	VIS	14.247	14.203	14.462	14.468	0.024	0.323
WRI	CIG	2.857	2.909	3.054	3.243	0.174	0.445
WRI	VIS	3.687	3.697	3.966	3.997	0.122	0.407

4. Subject to the stated constraints, the merits and demerits are determined so as to maximize the statistic defined by the t-test of the difference between the average scores of perfect forecasts and random climatological forecasts.

5. The perfect forecast receives a population average score of unity. (However, the sample average value need not equal unity.)

The merits and demerits for each predictand are exhibited in Table A-7. The maximum average scores obtainable from an informationless pure strategy are shown in Table A-8.

TABLE A-8
MAXIMUM AVERAGE SCORES OF
INFORMATIONLESS PURE STRATEGY

Sta	Elem	Max av score
ACY	CIG	0.108
ACY	VIS	0.043
CEF	CIG	0.148
CEF	VIS	0.085
DCA	CIG	0.125
DCA	VIS	0.158
IDL	CIG	0.133
IDL	VIS	0.123
OFF	CIG	0.162
OFF	VIS	0.096
RND	CIG	0.062
RND	VIS	0.061
WRI	CIG	0.131
WRI	VIS	0.071

A sample layout of the scoring scheme, indicating how the merits and demerits (Table A-7) should be interpreted, is shown in Table A-9. This illustration exhibits the scoring scheme for Westover ceiling.

A feature of the merits that might be puzzling is that, in classes 1 through 4 of a given predictand, the merits are slightly greater for the more probable classes. The merits and demerits are the consequences of several mathematical conditions. The main effect of these conditions, on the kind of probability distribution we are dealing with, is

TABLE A-9
EXAMPLE OF SCORING SCHEME
FOR CEILING AT WESTOVER AFB

Class		Forecast 1	2	3	4	5
Observed	1	2.357	0	0	0	-0.535
	2	0	2.381	0	0	-0.535
	3	0	0	2.500	0	-0.535
	4	0	0	0	2.766	-0.535
	5	-0.535	-0.535	-0.535	-0.535	0.246

to make the merits nearly uniform in classes 1 through 4; but a minor effect is to make them increase slightly with increasing probability of the classes. The scores have been designed for comparing skill, and, among the first four categories, those that occur more often afford greater opportunity for making comparisons. As it turns out, the merits in classes 1 through 4 are roughly equal to the reciprocal of the probability of the nonoccurrence of class 5. Broadly speaking, the general operation of the scoring scheme is to measure skill in distinguishing between the occurrence and nonoccurrence of class 5 and at the same time to compare hits in classes 1 through 4.